Optimization of User Query for Improving Document Retrieval Performance

Nidhi Bhandari
IET, DAVV, Khandwa Road,
Indore, India

Rachna Navalakhe, PhD
SGSITS, 23, Park Road, Indore,
India

G.L. Prajapati, PhD
IET, DAVV, Khandwa Road,
Indore, India

ABSTRACT

The unstructured data processing and finding accurate information from IR models is a complex task. The classical techniques use different concepts for improving IR models such as categorization, classification and many more. This paper reviews different document retrieval techniques first and then an extension on previously introduced version is provided. Similar to the traditional model, this technique first pre-process data and extract features. After that the retrieved features are organized in a tuple. These tuples are further used with fuzzy c means algorithm to cluster their domain according to their features. This process reduces the time of proposed search model. In addition to that, for preventing inappropriate query submission, the new query generation and optimization is also proposed in this work. The results with the different dataset shows the proposed IR model improve the performance in terms of efficiency and accuracy.

Keywords

Text mining, Query optimization, Semantic knowledge, Information retrieval, c-means clustering

1. INTRODUCTION

The text mining techniques are data mining algorithms employed for recovering user query based outcomes [1]. In this work an extension of previously introduced IR model is provided. In order to improve the existing model a semantic query optimization technique is used. This helps to improve the probability to extract maximum information according to user query [2]. The IR model contains three key components (1) Accept user query from user (2) Query processing (3) Generation of outcomes suitable to the user query. The proposed work is effective for optimizing the user query and improving the search time reducing the search space. Basically, lack of user query keyword or inappropriate keyword selection, search system produces false results [3]. Therefore, it is required to optimize user input queries for finding accurate user required content. Additionally, for reducing down the time complexity of system, there is need to organize the extracted text features according to their domain [4]. The proposed work contributes the following:

1. Optimization of user query
2. Reducing the system learning time
3. Improving accuracy of search results.

This paper is intended to improve the performance of the existing IR model by incorporating the query optimization and the clustering of the extracted features to reduce down the cost of search. Therefore, the proposed model with the modification is reported in this paper and a comparative performance study is involved for justifying the proposed work. This section provides overview of the proposed improvement on existing model; next section explains the recent supporting contributions by the different authors and researchers. In next section, the proposed information retrieval technique is explained and finally the results are calculated and their performance is compared. At last the future directions of the proposed work are described.

2. IDENTIFY, RESEARCH AND COLLECT IDEA

This section offers study about recent literature and essential contributions to understand the working text mining and their techniques for improvements.

The topic modeling enable other applications i.e. search, information browsing, and pattern mining. Long Chen et al [5] proposed a semantic graph based topic modeling for structuring text streams. Model assimilates topic mining and time synchronization for handling lexical gap. For sources asynchronization problem, local semantic graphs are engaged. Similarly, Zhangjie Fu et al [6] propose a content-aware search scheme with semantic search. First, introduced conceptual graphs (CGs) for knowledge representation. Then, presented two schemes. To conduct calculation, CGs are transferred into linear form and map. Second, employmentof multi-keyword ranked search over encrypted data. Yan Chen et al [7] used a two-step process. Then, a summary is generated from micro blogs. This helps users to understand the possible understandings of retrieved results. The approach makes use of automatically learned information. First incorporated this information with a semi-supervised probabilistic graphical model and this helps to achieve significant classification performance.

Julien Ah-Pine et al [8] is interested in storehouses of image/text objects and study multimodal information fusion. Authors focus on graph based methods. Observed two methods: cross-media similarities and random walk based scores. And propose a graph based framework to involve two approaches. Junjie Cai et al [9] propose semantic attributes for image search re-ranking. A hyper-graph is used to model relationship between images by visual features and attribute features. This ranking is performed to order the images. Kazuo Aoyama et al [10] presented fast zero-resource spoken term detection (STD), by using hierarchical graph-based similarity search method (HGSS). HGSS is an improved graph-based similarity search method (GSS). A search algorithm for the hierarchical k-DR graph consumes the cluster structure to reduce search space. A vertex and an edge in hierarchical graph correspond to a Gaussian mixture model (GMM) and the relationship between a pair of GMM posterior gram segments, which is measured by dynamic time warping.

The retrieval function governs to what extent some information is relevant to a user query. Most retrieval functions have “free parameters” affecting effectiveness. Choosing the best values for such parameters is important. Alberto Costa et al [11] propose to regulate free parameter values by solving an optimization problem aimed at maximizing retrieval effectiveness. Authors engaged the black-box optimization, a simple grid-search and techniques such as line search and surrogate model. Results not only provide useful insight, but also provide efficiency. The objective of Saruladha Krishnamurthy et al [12] is to provide an insight into the information retrieval definitions to process models. The IR models have not only used for search it also supported cross lingual translation. This also outlines the CLIR process. The tools used for experiment and research are also discussed. This is organized as summary to the concepts of IR. Description of IR process, models, role of external sources, like ontologies is also included. Finally, it provides overview of CLIR and the tools used.

Modern search engines aggregate results from different verticals: webpages, images, video, etc., these results are heterogeneous in nature. This directly challenges the “ranked list” formulation. Therefore, finding proper presentation for a gallery of heterogeneous results is critical. Yue Wang et al [13] proposed a framework that learns the finest page presentation onto search result page. Page presentation is defined as the strategy to present a set of items, more expressive than a ranked list. It specifies positions, image sizes, text fonts etc. The presentation is content-aware, i.e. personalized to specific queries and returned results. Accurately answering of queries in scientific case and finding articles in literature requires capturing many latent aspects of information needs. Proper representation of query analysis is made to identify query concepts and query transformation. Saeid Balaneshin-kordan et al [14] proposed a method for representing domain specific queries based on weighted unigram, bigram and multi-term concepts as well as extracted from the top retrieved documents. The paper proposed a graduated non-convexity optimization framework, to allow unify query analysis and expansion depending on their type and source. Experiments indicate that applying this method results in improvement of retrieval accuracy.

Evaluation of IR systems focuses on a single measure that models the utility. Such measure usually combines a behavior based rank with a notion of document utility. However, individual users relevance such as sincerity, reputability or readability strongly impact on utility. For different needs the utility can be a different because of the focus on single metrics. Joost van Doorn et al [15] proposed to moderate this by viewing multiple relevance criteria and learning a set of rankers. The paper modelled document utility within a gain-based evaluation as a weighted combination. Using the learned set, it is able to make decision on the values of rankers and preference. Author showed that there is different available trade-offs between relevance criteria. Similarity measures define similarity between two or more documents. The retrieved documents are ranked on the similarity of content. Manoj Chahal et al [16] recovered information with the help of Jaccard similarity coefficient and with the help of Genetic Algorithm. Due to exploring and exploiting nature of Genetic Algorithm, it gives optimal result. GA use Jaccard similarity coefficient to calculate similarity between documents. Value of similarity function lies between 0 &1 shows the probability of similarity between the documents.

3. PROPOSED MODEL

The rising technology is responsible for improvements on the existing systems. This section provides the detailed discussion about the proposed IR model and the improvements made on existing system.

3.1 System overview

In this work the main aim is to improve the previously introduced information retrieval model. In order to enhance traditional IR issues. (1) Large running time: Due to large amount of data in database and lake of effective techniques a significant amount of time required to locate information. The involvement of clustering technique for learning data organization improves the search time. (2) Large domain of documents: the documents available in database are not necessary to available in a similar in category and contents. That increases search space. (3) Selection of ineffective keywords: most of the time users utilize irrelevant keywords for finding the required information. Thus need to optimize search query keywords. Therefore, to improve the existing IR system the key modifications are made on the basis of user query optimization and the domain categorization.

3.2 Methodology

The proposed improved IR model is demonstrated in two major parts i.e. training and testing or information retrieval. The training process of the proposed technique is defined in figure 1

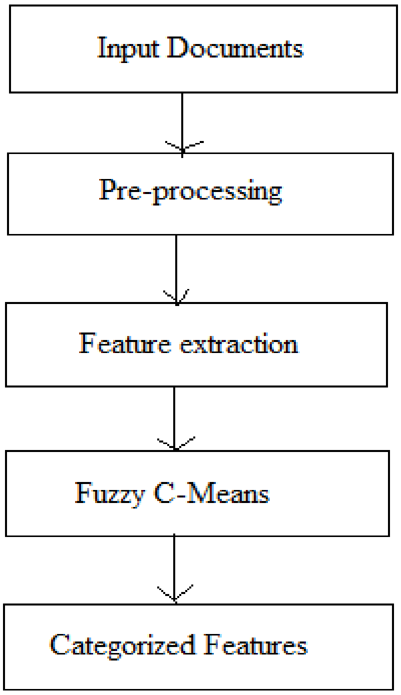

Figure 1 Training model

Input document: The IR techniques are contains data storage in an unstructured format. This storage contains documents in raw format. Additionally after completing the search process the documents are produced from this storage.

Data pre-processing: The unstructured data is complex form of data storage. It contains a significant amount of noise and unwanted contents. Therefore this step is used to improve the quality of data and reduce the noisy content. In this work two data pre-processing techniques are adopted (1) removal of stop words (2) the removal of special characters. Using this step, not only the documents data is reduced, memory space is also reduced to perform efficient search.

Feature selection: The unstructured documents length is different from each other. Therefore comparing the user query to all the documents are time consuming. Thus the advantage of the feature selection is to reduce contents for comparing them. To calculate the features the TF-IDF is used. The TF is calculated using following equation.

$$TF = \frac{total\ times\ a\ word\ appeared\ in\ document}{total\ words\ available\ in\ document}$$

And IDF using,

$$IDF = \log_e\left(\frac{total\ number\ of\ documents}{number\ of\ documents\ with\ the\ traget\ term}\right)$$

Using these two equations the weights are calculated as:

$$Weight\ W = TF * IDF$$

This weight highlights the important tokens for a document. Additionally to regular size of feature vector the 30 is created as maximum vector size.

Fuzzy c means: The fuzzy c means is a clustering algorithm which works on the basis of membership function. That is different from the partition based clustering algorithms. In this algorithm the single database object can belongs to multiple clusters. The fuzzy c means clustering is defined in the following manner. The fuzzy C-mean clustering involves an objective function:

$$J_m = \sum_{i=1}^{M}\sum_{j=1}^{C} u_{ij}^{m} \left\|x_i - c_j\right\|^2$$

Where, m is a real number greater than 1. $u_{ij}^{m}$is the degree of membership. x_i Is ith element of datac_j is centroid.

Therefore, in order to categorize a vector in two or more categories we need to select similar data instances randomly as centroid as c_j for example for two clusters j = 1, 2. Additionally thex_i is the instances (feature vector) available. Thus feature vector V can be defined as $V = \{x_1, x_2, \dots, x_o\}$. According to the requirement the objective function can be rewritten as:

$$J_O = \sum_{i=1}^{O}\sum_{j=1}^{2} u_{ij}^{O} \left\|x_i - c_j\right\|^2$$

Now need to compute degree of membership which is used for partitioning the data, the degree of membership u_{ij}^{O} is computed as:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{o-1}}}$$

After first phase of clustering that is essential for update the previous centroid. The new centroids are calculated using the following function.

$$c_j = \frac{\sum_{i=1}^{o} u_{ij}^{o} * x_i}{\sum_{i=1}^{o} u_{ij}^{O}}$$

The optimization process can take a significant time thus a termination condition is required. The algorithm stops working when the degree of membership in the step k and k + 1 remains constant or it is tends to 0 then the termination condition reached.

$$\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{k}\right|\right\} < \varepsilon$$

The clustering group the extracted data features according to their content similarity. Thus the resultant of clustering is well organized list of features which can be defined using the following tuple.

$$F =< F_n, k_{1,2,\dots n}, C >$$

Where F is defining the feature set, F_n is the file name or index, $k_{1,2,\dots n}$ is the list of feature keywords and C is the class name or cluster number where the file available.

The initial modification is performed on data organization or during storage of new information in database. Now the effort is made to improve the user query.

User query: The user input query is used to finding the information from the database. User query is also vectored by tokenizing the query keywords. In this context an additional database is prepared which contains the keywords and the relevant synonyms for the given word. Using this database the initial user query is reformed. In this context the following algorithm is used as described in table 3.1.

Table 1 Query enhancement algorithm

```
Input: user query $Q = \{q_1, q_2, \dots, q_k\}$, synonyms database $SDB_n$
Output: multiple queries $Q_n$

Process:
1. $Q_k = ReadUserQuery(Q)$
2. $for(i = 1; i \le k; i++)$
    a. $for(j = 1; j < n; j++)$
        i. $if\left(SDB_j.contains(q_i)\right)$
            1. $q_i = q_i.replace(q_j)$
            2. $Q.Append(q_i)$
        ii. $endif$
    b. $endfor$
3. End for
4. Return Q
```

The algorithm described in Table 1shows the process to enhance the user input query. Most of the time, due to selection of inappropriate keywords the required information is not retrieved , as the data base contains the similar semantic word but not exactly the same. Thus, this modification produces the different search query to increase the chances to find the target data from the database. After improvement in user query, the search is performed using the developed user queries and data base.

Search process: The search process is not much changed from the previously introduced process of search method. Therefore, the k-NN (k-nearest neighbor) algorithm is applied for searching. The overview of k-NN algorithm is provided in previous work. The extension of the previously provided technique is given here. The table 2 shows the steps of the proposed search algorithm.

Table 2 k-NN based technique

Input: feature database F_o, user input query Q_n

Output: search results R_m

```
Process:
1.  $For(i = 1; i \leq n; i++)$
    a.  $temp = Q_i$
    b.  $for(j = 1; j \leq o; j++)$
        i.   $DataVec = F_j$
        ii.  $D(temp, DataVec) = \sqrt{temp^2 - DataVec^2}$
        iii. $if(D(temp, DataVec) \leq 0.25)$
             1.  $count++$
             2.  $R_{count}.Add(dataVec.FileName)$
        iv.  $endif$
    c.  $endfor$
2.  $Endfor$
```

The mentioned algorithm table 2 demonstrates the working process of the proposed search algorithm. That algorithm finds the distance between each generated query string and available data in database features. When the distance between two instances of data remains less than 0.25 then it is added to the results. After implementing the entire process the results are computed for finding performance of the proposed system. In addition of that to demonstrate the effectiveness of the proposed work the two classically introduced models are compared in this work.

4. RESULTS ANALYSIS

This section provides discussion about evaluation of proposed system. Therefore, different performance parameters are computed and on observation basis line graphs are represented.

4.1 Precision

Precision in the application of information retrieval system can be defined by amount of relevant data retrieved for search query. That can be evaluated by the following formula:

$$precision = \frac{relevantdocument \cap retrieveddocument}{retrieveddocument}$$

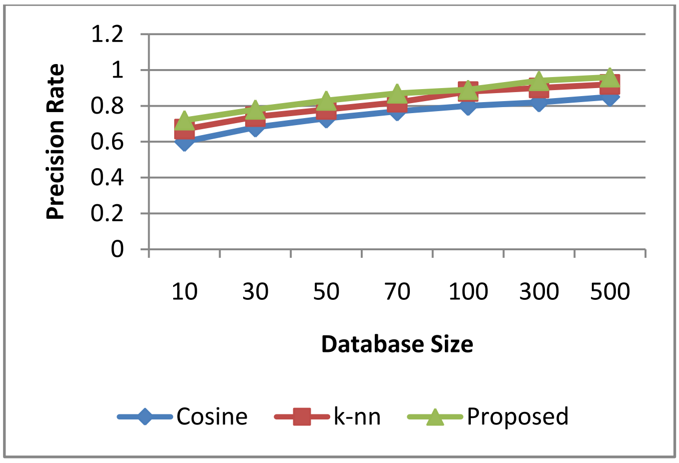

Figure 2 Precision Rate

The precision of an algorithm also recognized as the accuracy of an algorithm. The precision of the implemented three algorithms are demonstrated using line graph 2, the X axis of this line graph shows the number of data files available in database and the Y axis shows the precision rate of algorithm. The proposed technique is an extension of previously proposed k-NN based search system or information retrieval technique. According to the results the accuracy of all the approaches enhances with the size of data, but the modified technique is providing more precise results as compared to previously introduced k-NN based technique. That is become feasible due to improvement of query processing strategy in previous version of the IR system.

4.2 Recall

Recall of an information retrieval system is the amount of data that are extracted during the search is relevant to the user query. That can be measured using the following formula:

$$recall = \frac{RelevantDocuments \cap RetrievedDocument}{RelevantDocuments}$$

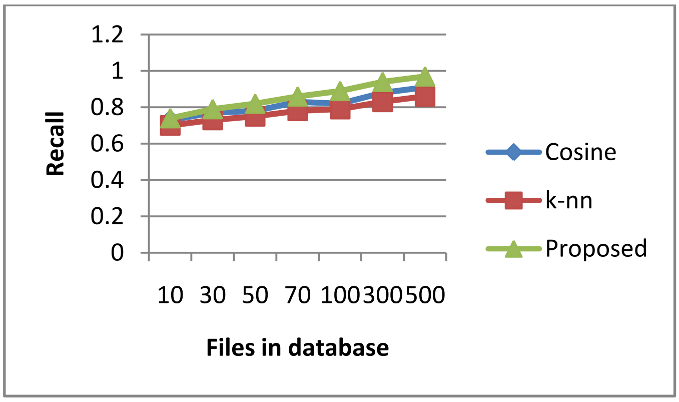

Figure 3 Recall rate

The recall of the extended k-NN based information retrieval model is given in figure 3. The line graph shows number of files in database in X axis and the Y axis contains the recall rate of the implemented systems. The blue line in this line graph shows the cosine simililarity based technique and red line shows the traditional k-NN based technique and the green line shows the extended technique in this paper. The recall of the proposed technique shows improved outcomes as compared to previously offered technique. Thus proposed technique is efficient and accurate.

4.3 F-measures

That measure and combines precision and recall in terms of harmonic mean of precision and recall rate of the obtained results, that can also be termed F-measure or balanced F-score:

$$F - measure = 2.\frac{precision * recall}{precision + recall}$$

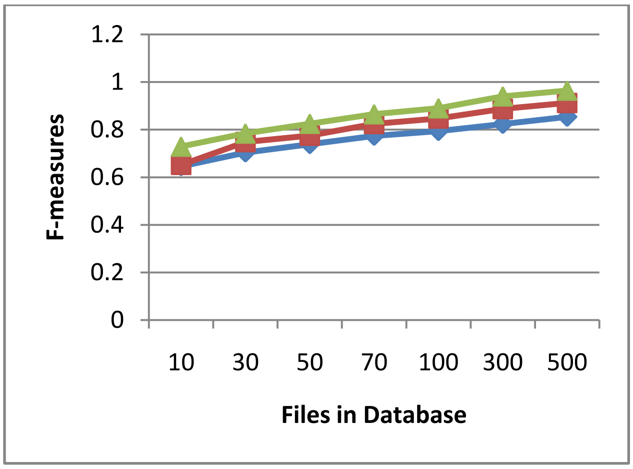

Figure 4 f-measures

The traditional techniques namely cosine similarity based IR technique, previously proposed k-NN based IR model and proposed extended k-NN based models are compared here in terms of f-score. The f-score shows the harmonic means of the IR systems. Thus it is also used to represent the tread off between precision and recall. According to the observed performance in figure 4 the proposed extended query processing based IR model perform much efficient and accurately as compared to our previously proposed model.

4.4 Memory usages

The main memory required for computing the outcomes of algorithm is known as memory utilization or space complexity of algorithm. That can be computed using the following formula.

$$memoryused = totalmemory - freememory$$

The memory usages or consumption of the introduced three algorithms are given in figure 5 .figure is a line graph and described in 2D, the X axis of the diagram shows the amount of file reside in database and the Y axis demonstrate the utilized memory for processing the user request. The memory usage of the algorithms is explained here in KB (kilobytes). According to demonstrated results of all three algorithms the proposed technique consumes utilized mean memory with respect to other two previously introduced algorithms. Therefore the proposed algorithm is memory efficient and utilizes limited memory resources.

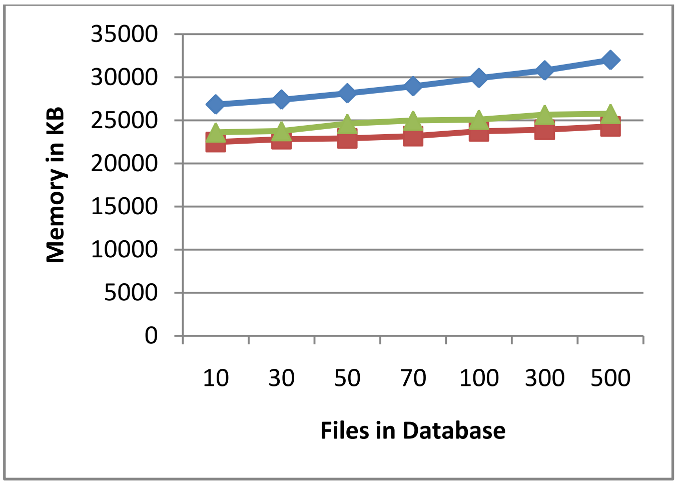

Figure 5 memory consumption

4.5 Time consumption

The time required for processing the user query in order to find relevant data from data base is termed here as time consumption. That is also known as the time requirement or time complexity. The time requirement of an algorithm is calculated using the following eq.:

$$timeconsumption = finishingtime - initiationtime$$

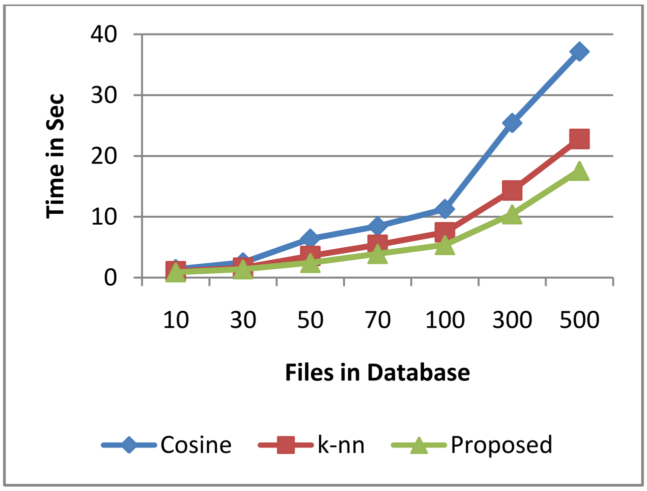

Figure 6 time consumption

The time consumption of the classically implemented and proposed technique with modifications is described in this section. The performance of the implemented IR techniques is represented using line graph in figure 6. The blue line of diagram shows the cosine similarity based approach, red line shows the k-NN based algorithms performance and the green line shows the proposed modified algorithm's performance. According to the demonstrated performance the proposed modified technique is efficient as compared to the previously given algorithm and the cosine simililarity based technique. Thus according to the all the evaluated parameters the proposed technique enhances the performance of previously introduced approach in terms of accuracy as well as the time and space complexity.

5. CONCLUSION

The proposed work is motivated for exploring the domain of information retrieval (IR). Therefore first the recently available development in IR is investigated. According to the obtained conclusion in literature we observed that the enhancement in user query and optimization in query keywords can improve the visibility of actual search results. By using this hypothesis the proposed work modifies the previously proposed technique of information retrieval for text data base. In order to extend the proposed text IR the efforts are concentrated on improving the user query inputs. Thus a query processing technique is proposed in this work, with the previously introduced model. In addition of that to reduce the search space of the proposed algorithm the FCM (fuzzy c means) clustering algorithm is used on database for learning with the database features and identification of document's domain. That technique is employed to reduce the cost of search in terms of time taken of the algorithm. The implementation of the required data model is performed in JAVA technology. The results measured about the performance of the system indicate the proposed extension in the previously proposed technique enhances the performance of the IR system. The implemented system is efficient and accurate but the possibility of improvement always remains. Thus in near future the following work is proposed for extension.

1. Implementation of indexing and ranking of the search results to improve visibility in search outcomes.
2. Implementing the directional graph model for improving the search results accuracy.

6. REFERENCES

[1] J. Han, M. Kamber, J. Pei, "Data Mining Concepts and Techniques", http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[2] H. Wang, Q. Zhang, & J. Yuan, "Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach", 2169-3536, 2017 IEEE

[3] S. Bergamaschi, E. Domnor, F. Guerra, M. Orsini, R. T. Lado, Y. Velegrakis, "Keymantic: Semantic Keyword-based Searching in Data Integration Systems", Proceedings of the VLDB Endowment, Vol. 3, No. 2, Copyright 2010 VLDB ACM

[4] G. Kumaran and J. Allan, "Simple Questions to Improve Pseudo-Relevance Feedback Results", Copyright is held by the author/owner(s), SIGIR'06, August 6–10, 2006, Seattle, Washington, USA ACM

[5] L. Chen, J. M. Jose, H. Yu, F. Yuan, "A Semantic Graph-Based Approach for Mining Common Topics from Multiple Asynchronous Text Streams", c 2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License. WWW 2017, April 3–7, 2017, Perth, Australia. ACM 978-1-4503-4913-0/17/04

[6] Z. Fu, F. Huang, K. Ren, J. Weng, and C. Wang, "Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data", IEEE Transactions on Information Forensics and Security, Vol. 12, No. 8, August 2017

[7] Y. Chen, X. Zhang, Z. Li, J. P. Ng, "Search engine reinforced semi-supervised classification and graph-based summarization of microblogs", Neurocomputing 152 (2015) 274–286

[8] J. A. Pine, G. Csurka, S. Clinchant, "Unsupervised Visual and Textual Information Fusion in CBMIR using Graph based Methods", ACM Transactions on Information Systems, Vol. , No. , 20, Pages 1–0??.

[9] J. Cai, Z. J. Zha, M. Wang, S. Zhang, and Q. Tian, "An Attribute-assisted Reranking Model for Web Image Search", IEEE Transactions on Image Processing, Vol. X, No. XX, Month Year

[10] K. Aoyama, A. Ogawa, T. Hattori, T. Hori, and A. Nakamura, "Zero-Resource Spoken Term Detection Using Hierarchical Graph-Based Similarity Search", 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)

[11] A. Costa, E. D. Buccio, M. Melucci, G. Nannicini, "Efficient Parameter Estimation for Information Retrieval using Black-box Optimization", IEEE Transactions on Knowledge and Data Engineering (Volume: 30 , Issue: 7 , July 1 2018)

[12] S. Krishnamurthy, Akila V, "Information Retrieval Models: Trends and Techniques", Copyright © 2018, IGI Global.

[13] Y. Wang, D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, Q. Mei, "Beyond Ranking: Optimizing Whole-Page Presentation", WSDM'16, February 22–25, 2016, San Francisco, CA, USA. c 2016 ACM. ISBN 978-1-4503-3716-8/16/02

[14] S. Balaneshin-kordan, A. Kotov, "Optimization Method for Weighting Explicit and Latent Concepts in Clinical Decision Support Queries", ICTIR '16, September 12-16, 2016, Newark, DE, USA c 2016 ACM. ISBN 978-1-4503-4497-5/16/09

[15] J. v. Doorn, D. Odijk, D. M. Roijers, M. d. Rijke, "Balancing Relevance Criteria through Multi-Objective Optimization", SIGIR '16, July 17 - 21, 2016, Pisa, Italy c 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4069-4/16/07

[16] M. Chahal, "Information Retrieval using Jaccard Similarity Coefficient", International Journal of Computer Trends and Technology (IJCTT) – Volume 36 Number 3 - June 2016.